View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Revistes Catalanes amb Accés Obert

revista.tradumàtica
tecnologies de la traducció

484-492

Linguistic Quality Review: A Case Study

Willem Stoeller PMP
International Consulting, LLC

ABSTRACT

This article describes the approach to quality measurement of human translation and measuring translator performance at VMware, a virtualization software company. The article focuses on the creation of a review environment using error typology combined with sampling and performance trends.

Keywords: Error typology, linguistic quality, sampling, translator performance, quality review.

RESUM *(Revisió de qualitat lingüística: estudi de cas)*

Aquest article descriu com acostar-se a la mesura de la qualitat en traducció humana, així com al mesurament del rendiment del traductor en VMware, una companyia de virtualització de programari. L'article se centra en la creació d'un entorn de revisió que utilitza una tipologia d'errors combinada amb les tendències de mostreig i de rendiment.

Paraules clau: tipologia d'errors, qualitat lingüística, mostreig, rendiment del traductor, revisió de qualitat.

RESUMEN *(Revisión de calidad lingüística: estudio de caso)*

Este artículo describe como acercarse a la medición de la calidad en la traducción humana, así como a la medición del rendimiento del traductor en VMware, una compañía de virtualización de software. El artículo se centra en la creación de un entorno de revisión que utiliza una tipología de errores combinada con las tendencias de muestreo y de rendimiento.

Palabras clave: tipología de errores, calidad lingüística, mostreo, rendimiento del traductor, revisión de calidad.

1. Introduction

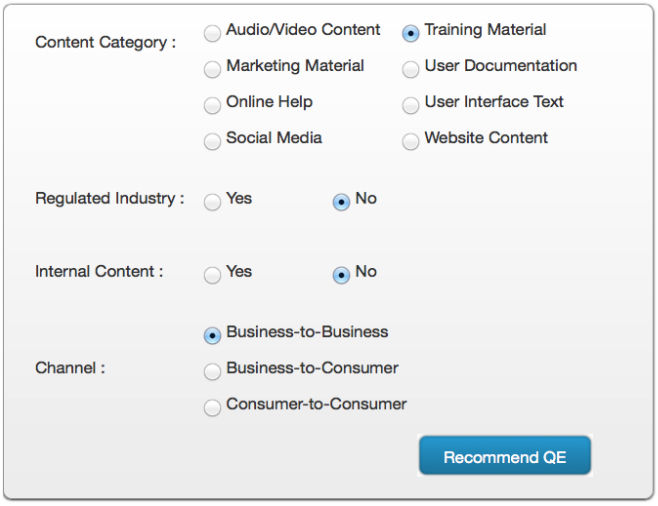
Since May 2013 I have been engaged as a localization consultant with the VMware department responsible for the localization of marketing, web and educational content. One of my priorities is to assist this department with the implementation of linguistic quality evaluation processes. The quality processes in place at that time consisted of full review of all translated content by a Localization Service Provider (Review Vendor) other than the LSP responsible for the original translation. Not only did the Review Vendor inspect hundred percent of all translated content, they also made all changes deemed necessary. However formal quality evaluation was only done sporadically using a very simple error typology. VMware felt that this approach was too costly, time consuming and subjective, and certainly not scalable.

As a long time member and representative of TAUS I was very familiar with the Dynamic Quality Framework (DQF) developed by the TAUS members. Thus I turned to the content profiling tool on the TAUS DQF website. This tool recommends best practices in quality

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

evaluation based on the content type and communication channels used.

Figure 1: DQF (TAUS) Content profiling for marketing content



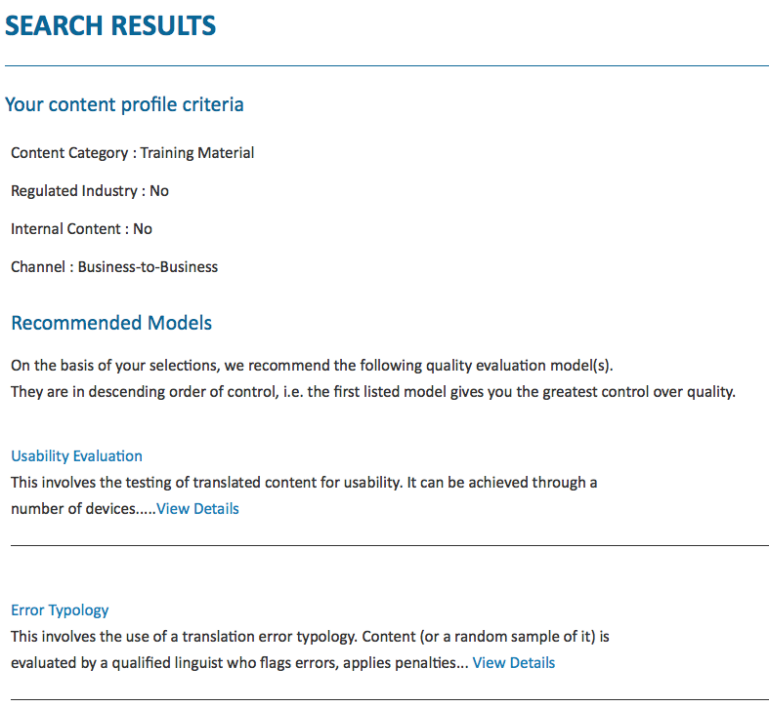


Figure 2: DQF (TAUS) Content profiling for training content

The recommendations of the content profiling tool confirmed my own experience, therefore I suggested using a customized error typology as a quality evaluation approach for marketing, web and training content. Although usability evaluation was the DQF recommended technique for training material, VMware considered this approach too costly and time-consuming for translated content. Henceforth the starting point for the development of quality review

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

processes at VMware was the error typology template in the DQF (TAUS) Knowledge database.

According to Eduardo D'Antonio (Director of Globalization Operations for VMware), "The TAUS DQF tools and the support of its co-creators provided valuable input into our new quality evaluation processes. In addition, TAUS representative Willem Stoeller was instrumental in designing and implementing those quality evaluation processes."

2. Main body of article

2.1 The quality evaluation process

Traditional quality inspection and Six Sigma have taught us that producing consistently defect-free translation is not possible, therefore it is important to set an objective measure of quality. One approach to such an objective measure of quality is scoring translations using an error typology with a predefined tolerance for errors. At VMware translated content is scored using an error typology derived from TAUS' DQF. The TAUS Linguistic Quality Evaluation (LQE) scorecard uses four error categories (Accuracy, linguistic, terminology and style) and four levels of severity for each error category. Each combination of error category/severity is assigned a number of penalty points (the weight of the error type/severity combination) except for severity level 4. The latter severity level is used to indicate preferential changes and does not carry any penalty points. In order to pass an LQE review the translation cannot have more penalty points than an error threshold (normalized per/to 1,000 words).

Not all content types and content usages are equal; the TAUS' DQF states that the quality needed for a translation depends on the content type and communication channel (sender and receiver of the communication). Based on the potential impact of translation errors it is possible to define different levels of quality for different content type/communication channel combinations.

VMware identified three levels of quality risk (high, normal and low), they also derived three different LQE scorecards from the TAUS' DQF template: marketing, technical and adaptation (Transcreation). The only difference between these three scorecards is the number of penalty points assigned to each error category/severity combination. The threshold and all calculations are the same for all three scorecards. The content owners/stakeholders assigned each VMware content type a quality level and scorecard to be used for evaluation. All VMware training content is evaluated using the technical scorecard. VMware Marketing and web content are evaluated using the marketing, technical or adaptation scorecard.

The original TAUS template only had two outcomes: fail or pass. In VMware's situation it was felt that a quality metric was necessary. This was achieved by means of a pass score and an achieved score, both expressed as a percentage. See below for the calculations:

Metric	Definition
Threshold	The threshold is the maximum allowed penalty points to pass*word count /1,000
pass Score	The pass score is a percentage to indicate the quality level acceptable for a translation and is calculated as follows: pass score = 1-(threshold/(word count)

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

Metric	Definition
Achieved Score	The achieved score (called Current Score in figure 3) is a percentage to indicate the quality level of a translation and it is calculated as follows: 1. If the total penalty points assigned to the translation = 0, the achieved score is 100% 2. If the total penalty points assigned to the translation > 0, the achieved score = 1 – ((total penalty points assigned to the translation) /word count)

Figure 3: Definitions

Additionally some calculation exceptions were defined to deal with boundary conditions such as very small translations and critical errors. Very small translations (no more than 250 words) are reviewed and corrected by the reviewer but not scored (to save time). Translations with one or more severe errors (severity = 1) are automatically forced to fail.

Job Info

Review date	18-Mar-14	Test number	1
Source Language	EN-US	Program Type	
Target language	ES-LA	Translation Provider	
LQE Provider		VMGlobal Dispatch ID	12346
Reviewer's name		VMGlobal Project Name	Another Project Name
Word Count	500	Pass score	98.5%
Threshold	7.5	Current Score	100.0%
Quality Level	High	Minutes for Review	48
Sampling Percentage	100.00%	LQE Job ID	54320
		TMS Job ID	10001

Figure 3: VMware completed scorecard results

If the review results in a fail, necessary changes will be made to the translation such that the final reviewed and edited content always meets the pass criteria defined in the scorecards: For full reviews the reviewers always implement changes, even in the case of a pass in order to save time. For sampled reviews with a pass, the reviewer will also implement changes. For sampled reviews with a fail, the translator will rework the entire content, not only the reported errors.

The determination of an achieved score and pass score for each translation made it easy to calculate and plot average translator by translation vendor over a user-selected time period as shown below in figure 4. VMware also tracks performance score from quarter to quarter for each translation vendor. Each vendor need to have obtained an average achieved score (Score in Figure 4) that is larger than the average pass score (Target in figure 4) for at least one quarter before that vendor's translated content can be reviewed on a sampling basis.

VMware reduces the variance between its reviewers as much as possible through training. It is cost prohibitive to use multiple reviewers in parallel for translation production with hundreds of projects each month (multiple parallel reviewers would allow to take reviewer agreement into account).

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

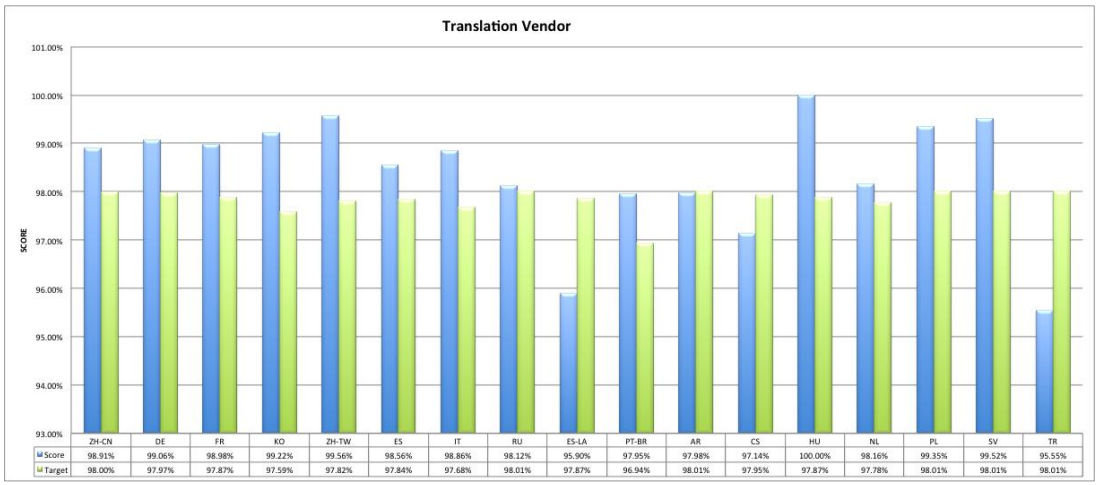


Figure 4: VMware translator report for the first quarter (source language is English (US), target language is as listed in graph

In addition the counts of fail and pass over a user-selected time period is tracked:

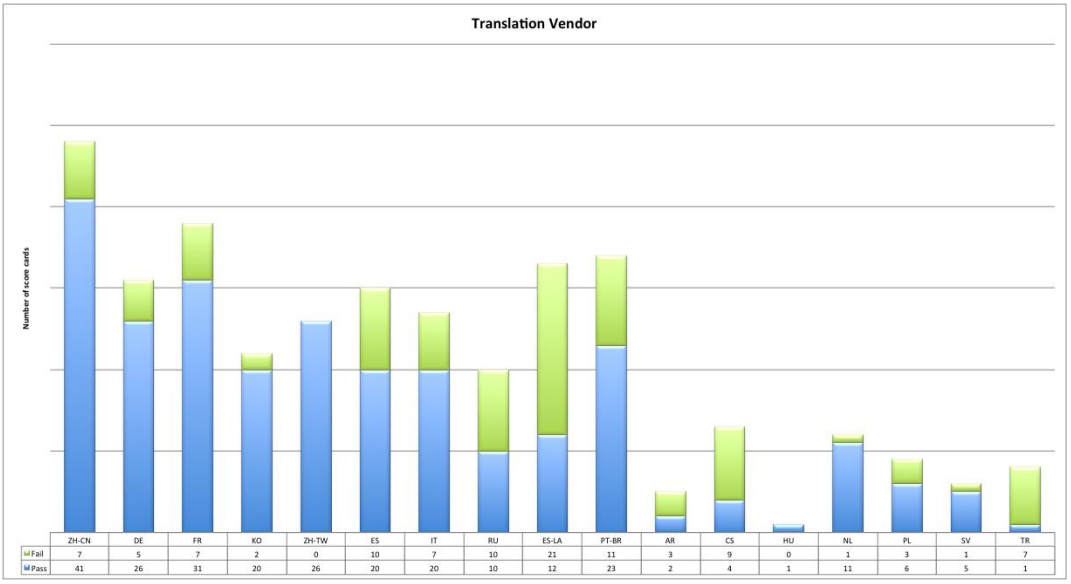


Figure 5: Counts of fail and pass in first quarter for different target languages

VMware, as is now common in the localization industry, holds Quarterly Business Reviews with their translation partners. In the Quarterly Business Reviews the performance reports are used to determine a single Key Performance Indicator and target for linguistic quality. This is done by not only averaging over the quarter but also averaging achieved and passes scores over all languages. (These results are included in the vendor's Balanced Scorecard).

All review scores and other related data elements are stored in a database, which VMware mines for information on potential process issues on both VMware and the vendor's side.

2.2. Sampling to reduce turnaround and cost

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

Many translation buyers who use an independent review vendor to score translated content, have concluded that it is too expensive and time consuming. DQF suggests a number of other quality evaluation methods that are less time intensive than using an error typology, but for the most common content types (web content, marketing, training, online help and user interfaces) the DQF's content profiler still recommends using an error typology.

Random Sampling

This is not a new problem or one specific to translation: product quality inspection, polling and medical research all share t same cost and time concerns. One solution is to select a representative subset of the entire translation and only review that subset. But he question then arises: "What can we say about the number of errors in the full translation based on the errors found in the subset?" The statistics of simple random sampling provides us with a possible answer:

Sample size	Total # of words	Confidence interval	Sampling error
20%	1,500	99.9%	8.5%
20%	5,000	99.9%	4.7%
20%	10,000	99.9%	3.3%
20%	20,000	99.9%	2.3%
20%	50,000	99.9%	1.5%
20%	100,000	99.9%	1.0%

Figure 7: Statistical sampling errors for simple random sampling

A few examples will illustrate the concept of the statistical sampling error.

Example 1: If we take a 20% random sample from a 1,500-word translated document and we find 3 errors, then we can deduct with 99.9% confidence that the entire document has 15 errors plus or minus 2 errors.

Example 2: If we take a 20% sample from a 5,000-word translated document and we find 5 errors, then we can deduct with 99.9% confidence that the entire document has 25 errors plus or minus 1 error.

Example 3: If we take a 20% sample from a 20,000-word translated document and we find 10 errors, then we can deduct within 99.9% confidence that the entire document has 50 errors plus or minus 1 error.

A word of warning here, the quality inspection profession moved on to Six Sigma, however that is a level of quality we cannot expect in translation industry any time soon. It would mean less than four errors per million words of translated content.

Systematic Sampling

Another approach to sampling is to select that subset of the translation where translation errors have the greatest impact. This is called systematic sampling in the TAUS Guidelines on Sampling report. However there is very little information available in the translation industry on how to achieve systematic sampling and no known tools available for this purpose at the time of writing. The manual selection of a systematic sample is usually left to the reviewer, based on their experience.

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

Sampling at VMware

VMware decided to use sampling for translation review as a way to reduce both the cost and the time needed for linguistic review. VMware applies three levels of quality each with their own linguistic quality review requirements:

1. Quality risk high: always review all of the translated content,
2. Quality risk normal: always review of a 20% sample only (Assuming the translation vendor meets the required performance levels described below),
3. Quality risk low: no review.

It is also important to consider the scope of the review: should the review include perfect matches (also known as in context matches) and exact matches? VMware's choice again depends on the quality level set for the translation:

1. Quality risk high: exclude perfect matches (in context matches) and include exact matches,
2. Quality risk normal: exclude perfect matches (in context matches) and exact matches,
3. Quality risk low: not applicable.

VMware introduced several additional rules:

1. Content with a word count of 250 or less is reviewed but not scored (in order to save time),
2. Content of quality risk normal with a word count of 1,000 or less is 100% reviewed and scored using the error typology described in Section 2.1,
3. In order for a particular translation vendor's translated content of quality risk normal to be reviewed on a 20% sampling basis, this translation vendor needs to have demonstrated a predefined level of performance for that particular language pair,
4. If the translation vendor, whose work is sampled for a particular language, does not meet the predefined performance requirements for sampled translation of quality level 2, then the review level goes back to full review instead of sampling.

To select a 20% random sample manually (using a random number generator) would be extremely time-consuming and also costly. Therefore VMware initially opted for a systematic sampling approach borrowed from another company: based on an agreed upon productivity rate of the reviewer, the time it takes to review 20% of the translated content is calculated (including extra time spend completing the scorecard). The reviewer is asked to spend that time reviewing/scoring manually selected segments from the translation. The validity of this approach depends on how the reviewer's performs this manual selection. For example if the reviewer uses the allotted time to review segments sequentially from the start of the document until the time is used up then the selected segments are not a representative subset of the entire translation, nor is the selection likely to be the subset with the highest risk for translation errors.

Therefore, to enhance the reliability of the sampling process, VMware is currently creating a review environment that will enable random selection of 20% of the words to be reviewed using the rules described above. In order to provide context for the reviewer the entire content will be shown in the review user interface, but those not selected will be locked. In addition a functionality allowing simple navigation to the next segment to be reviewed will be provided.

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)

2.3. A custom review environment

VMware started their new linguistic quality review processes with Excel based scorecards and mostly manual processes. The Excel based scorecards have been replaced with a browser based review environment that accepts source and translated data from any Translation Management System (TMS) in XLIFF format. The review and scoring can now take place online. The same review environment also can generate the performance reports discussed earlier in this article. The next stage will see random sampling integrated with the review environment.

3. Conclusion

VMware's implementation of an error typology derived from the DQF template has been successful, and project deadlines were met. It is too early yet to draw more precise conclusions regarding time gained and money saved. VMware intends to continually refine their linguistic quality review processes (for example by making the penalty points and threshold language specific and by using MQM (see reference below in section 4) for content specific error typology). In addition VMware will follow closely any new developments in TAUS' Dynamic Quality Framework, especially in the areas of systematic sampling and other review techniques such as readability and accuracy assessment.

4. References

The Dynamic Quality Framework website: https://evaluation.taus.net

Sample error calculators:

https://www.dssresearch.com/KnowledgeCenter/toolkitcalculators/sampleerrorcalculators.aspx

Six Sigma: https://en.wikipedia.org/wiki/Six_Sigma

Best Practices on sampling: https://evaluation.taus.net/resources-c/guidelines-c/best-practices-on-sampling

QT Launchpad (MQM): http://www.qt21.eu/launchpad/

Els continguts de la revista estan subjectes a una llicència Creative Commons (CC BY 3.0)